

Version dated: February 21, 2017

Generalising rate heterogeneity across sites in statistical phylogenetics

SARAH E. HEAPS¹, TOM M. W. NYE¹, RICHARD J. BOYS¹,
TOM A. WILLIAMS², SVETLANA CHERLIN³ AND T. MARTIN EMBLEY⁴

¹*School of Mathematics & Statistics, Newcastle University, Newcastle upon Tyne, NE1 7RU, U.K.*

²*School of Earth Sciences, University of Bristol, Bristol, BS8 1RJ, U.K.*

³*Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, NE1 3BZ, U.K.*

⁴*Institute for Cell and Molecular Biosciences, Newcastle University, Newcastle upon Tyne, NE2 4HH, U.K.*

Abstract

In phylogenetics, alignments of molecular sequence data for a collection of species are used to learn about their phylogeny – an evolutionary tree which places these species as leaves and ancestors as internal nodes. Sequence evolution on each branch of the tree is generally modelled using a continuous time Markov process, characterised by an instantaneous rate matrix. Early models assumed the same rate matrix governed substitutions at all sites of the alignment, ignoring the variation in evolutionary constraints. Substantial improvements in phylogenetic inference and model fit were achieved by augmenting these models with a set of multiplicative random effects that allowed different sites to evolve at different rates which scaled the baseline rate matrix. Motivated by this pioneering work, we consider an extension which allows quadratic, rather than linear, site-specific transformations of the baseline rate matrix.

We derive properties of the resulting process and show that when combined with a particular class of non-stationary models, we obtain one that allows sequence composition to vary across both sites of the alignment and taxa. Formulating the model in a Bayesian framework, a Markov chain Monte Carlo algorithm

for posterior inference is described. We consider two applications to alignments concerning the tree of life, fitting stationary and non-stationary models. In each case we compare inferences obtained under our site-specific quadratic transformation, with those under linear and site-homogeneous models.

(Keywords: Phylogenetics; Across site rate heterogeneity; Compositional heterogeneity; Tree of life.)

1 Introduction

In statistical phylogenetics, the goal is to learn about the evolutionary relationships amongst a collection of species, generally using DNA or protein sequence data. These relationships are represented through a rooted, bifurcating tree called a phylogeny. Substitutions in the molecular sequence alignment are typically modelled using continuous time Markov processes, parameterised through an instantaneous rate matrix. Early phylogenetic models were simplistic, generally assuming that the evolutionary process was in its stationary distribution and that substitutions at each site of the alignment could be described by the same underlying rate matrix. Under these models, the probability of change from one character state to another was therefore independent of both organismal lineage and the biochemical function of the nucleotide or amino acid in question. These simplifying assumptions were known to be false, but were made for the sake of mathematical convenience and computational tractability given the limited computing power for model fitting available at the time. In particular, it was already clear to early molecular evolutionists that rates of evolution vary as a result of the functional and structural constraints acting on a site: important sites evolve slowly because most mutations that arise at those sites are eliminated from the population by negative selection (Fitch and Markowitz 1970). Uzzell and Zorbin (1971) showed that the numbers of substitutions occurring at different sites could be modelled using a negative binomial distribution. Later, Yang (1993) incorporated the idea into statistical phylogenetics by allowing different sites to evolve at different rates. These rate parameters scaled the underlying Markov process rate matrix and were modelled as multiplicative random effects, with a unit mean gamma distribution.

Incorporation of across-site rate variation into standard, stationary substitution models has led to major improvements in model fit and to the accuracy of phylogenetic inference (Yang 1996). But there are other, pervasive features of molecular sequence data that these models do not accommodate. In particular,

nucleotide composition is believed to vary across both the sites of the alignment and the branches of the phylogenetic tree. For example, the GC-content of ribosomal DNA genes varies from 45-74% across the known diversity of cellular life (Cox et al. 2008), implying that the probabilities of each of the four nucleotides can change over time. These compositional shifts might reflect changing biases in DNA repair enzymes (Sueoka 1988) or, at least for genes encoding structural RNAs, adaptation to different growth temperatures (Galtier and Lobry 1997). As well as variation in sequence composition across taxa, there is also compositional variation observed among the different sites within an individual protein-coding sequence: due to functional constraints, most sites can tolerate only a limited, and typically biochemically homogeneous, subset of the twenty amino acids (Fitch and Markowitz 1970). The result is that, in addition to varying in evolutionary rate, sites can also differ in sequence composition. Neither across-branch nor across-site compositional variation are accommodated by standard stationary models, even after incorporating across-site rate variation. But as with heterogeneity in evolutionary rates, failure to account for variation in composition can lead to model misspecification and, therefore, serious phylogenetic error, as demonstrated by a number of empirical studies (Embley et al. 1993; Foster 2004; Lartillot et al. 2007; Philippe et al. 2011). Although a few models have been developed to jointly model both sources of heterogeneity (Blanquart and Lartillot 2008; Jayaswal et al. 2014), they have not been widely used in practice due to computational difficulties with model-fitting.

In a simple phylogenetic model, evolution at all sites is controlled by a single instantaneous rate matrix. The across-site rate variation model offers greater flexibility by allowing site-specific *linear* transformations of the baseline rate matrix. The improvement afforded by this simple modification served as motivation for this paper, and we describe a natural generalisation through a model that provides site-specific *quadratic* transformations of the baseline matrix. This allows qualitatively different patterns of transitions at different sites. Further, we demonstrate that when linear or quadratic across-site transformations are combined with a class of non-stationary Markov processes, we obtain computationally tractable models that allow sequence composition to vary both across branches of the tree and across sites of the alignment. The remainder of this paper is organised as follows. Section 2 introduces phylogenetic models of sequence evolution and the incorporation of multiplicative random effects to allow rate variation across sites. Section 3 describes our quadratic generalisation and its properties. In Section 4 we combine across-site linear and quadratic transformations with a general class of non-stationary substitution models and describe the properties of the resulting

Markov processes. Section 5 addresses the issue of inference for models incorporating our quadratic transformation. Specifically, we take a Bayesian approach to inference and describe the posterior distribution of interest and details of our numerical approach to model-fitting via Markov chain Monte Carlo sampling. In Section 6 we consider analyses of two biological data sets; the first involving a stationary model and the second, a non-stationary model. In each case we compare the performance of a site-homogeneous model with analogous models incorporating linear and quadratic across-site transformations of the baseline rate matrix. Finally, we summarise our conclusions in Section 7.

2 Phylogenetic models of sequence evolution

Denote by $y = (y_{i,j})$ an alignment of molecular sequence data where $y_{i,j} \in \Omega_K$ is the character at the j th site for taxon i and Ω_K is an alphabet with K characters, for example, the DNA alphabet with $\Omega_4 = \{A, G, C, T\}$. Denote the number of sites (columns) by M and the number of taxa (rows) by N and let $\mathbf{y}_j = (y_{1,j}, \dots, y_{N,j})^T$ be the j -th column in the alignment. Consider a rooted, bifurcating tree τ , with branch lengths ℓ , representing the evolutionary relationships amongst this collection of N taxa. For every site, phylogenetic models typically assume that evolution along each branch of the tree can be modelled using a continuous time Markov process $Y(t)$, characterised by an instantaneous rate matrix $Q = (q_{u,v})$ which has positive off-diagonal elements and rows that sum to zero. This matrix controls the dynamics of the substitution process through the matrix equation $P(\ell) = \{p_{u,v}(\ell)\} = \exp(\ell Q)$, where $p_{u,v}(\ell) = \Pr(Y(\ell) = v | Y(0) = u)$ for $u, v = 1, \dots, K$ is the probability of transitioning from character u to character v along a branch of length ℓ .

Standard phylogenetic models assume that the underlying continuous time Markov process is time reversible and in its stationary distribution $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K) \in \mathcal{S}_K$ where $\mathcal{S}_K = \{(x_1, \dots, x_K) : x_i \geq 0 \ \forall i, \sum x_i = 1\}$ denotes the K -dimensional simplex. Reversibility implies that $\pi_u p_{u,v}(\ell) = \pi_v p_{v,u}(\ell)$ for all u, v and allows the rate matrix to be represented in the form $Q = S\Pi$, where $\Pi = \text{diag}(\boldsymbol{\pi})$, and S is a symmetric matrix whose off-diagonal elements, $\rho_{u,v}$ with $\rho_{u,v} = \rho_{v,u}$, are termed *exchangeability* parameters. The latter determine the general propensity for change between the different pairs of characters. We define a rate matrix as reversible if it permits a parameterisation of this form. The most general reversible rate matrix, with $K(K-1)/2$ distinct exchangeabilities, characterises the general time-reversible (GTR) model.

Other commonly used substitution models are special cases. For example, the TN93 model is a special case for nucleotide data where $\rho_{C,T} = \rho_{T,C} = \rho_1$, $\rho_{G,A} = \rho_{A,G} = \rho_2$ and all other $\rho_{u,v}$ are equal to β . This simplification reduces the number of exchangeabilities from six to three but retains biological realism by allowing transversions (substitutions between a pyrimidine and a purine) and the two types of transitions (substitutions between pyrimidines and between purines) to occur at different rates, here β , ρ_1 and ρ_2 respectively.

Classically, the sites of the alignment y are assumed to evolve independently of each other and so the likelihood is given by

$$p(y|Q, \tau, \ell) = \prod_{j=1}^M \Pr(\mathbf{Y}_j = \mathbf{y}_j | Q, \tau, \ell).$$

In order to prevent compensatory rescaling of the branch lengths ℓ and the rate matrix Q in the transition matrix $P(\ell) = \exp(\ell Q)$ it is common to impose an identifiability constraint on the rate matrix, for example by fixing one of the exchangeability parameters $\rho_{u,v}$, $u \neq v$, to be equal to one. For instance, one can fix $\beta = 1$ in the TN93 model. This allows the remaining exchangeability parameters to be interpreted as relative rates of change.

2.1 Modelling rate heterogeneity across sites

It has long been recognised that selective pressures vary across sites due to their differing roles in the structure and function of the molecular sequence. This feature is typically captured by allowing each site j to evolve at its own rate $c_j > 0$ which scales the rate matrix Q linearly. To enable information to be shared between sites, the rates $\mathbf{c} = (c_1, \dots, c_M)^T$ are generally assumed follow a gamma distribution with unit mean. The likelihood can then be represented as

$$p(y|Q, \tau, \ell, \alpha) = \prod_{j=1}^M \int_0^\infty p(c_j | \alpha) \Pr(\mathbf{Y}_j = \mathbf{y}_j | Q_j, \tau, \ell) dc_j,$$

where

$$Q_j = c_j Q \tag{1}$$

and $p(c_j | \alpha)$ is the $\text{Gam}(\alpha, \alpha)$ density function evaluated at c_j . The single parameter α determines the manner and extent to which the scaling factors differ across sites. We refer to models in which a baseline rate matrix is transformed according

to (1) as *linear across site heterogeneity* (LASH) models. In order to simplify computation, the (continuous) gamma density $p(c_j|\alpha)$ is typically replaced by a discrete approximation with K_c categories, most often $K_c = 4$ (Yang 1994). In a Bayesian setting, this numerical integration strategy may seem less natural than using data augmentation during MCMC and sampling the c_j . However, the discretisation allows much more caching of intermediate likelihood calculations which can substantially speed up computational inference.

In this model, the rate matrix at each site is simply a linearly scaled version of some underlying base matrix Q . The transformation does not affect the theoretical stationary distribution, defined as the solution of $\pi Q = \mathbf{0}^T$, or, in the class of reversible models, the ratios of the exchangeability parameters. In the following section we generalise this model to allow the rate matrix at each site to be a *quadratic* function of the base matrix, which depends on the values of *two* parameters. This allows the patterns of substitution, as well as the overall substitution rate, to vary between sites.

3 Quadratic across site heterogeneity (QuASH) models

Consider a baseline rate matrix Q . At site j , the instantaneous rate matrix $Q_j = (q_{j,u,v})$ is given by

$$Q_j = c_j Q - c_j d_j Q^2 \quad (2)$$

where $c_j \in (0, \infty)$ and $d_j \in (l(Q), u(Q))$, which reduces to the simple LASH model when $d_j = 0$. We call any model in which a baseline rate matrix is transformed in this way a *quadratic across site heterogeneity* (QuASH) model. The limits $l(Q)$ and $u(Q)$ depend on Q and ensure that Q_j is a valid rate matrix, that is (i) all off-diagonal elements are positive: $q_{j,u,v} > 0, \forall u \neq v$; (ii) all row sums are zero: $\sum_v q_{j,u,v} = 0 \forall u$.

Property (ii) is automatically satisfied for any $d_j \in \mathbb{R}$. The proof is as follows. The (u, v) -th element of Q_j is given by

$$q_{j,u,v} = c_j \left(q_{u,v} - d_j \sum_w q_{u,w} q_{w,v} \right).$$

Therefore the sum of the elements on row u of Q_j is

$$\begin{aligned}\sum_v q_{j,u,v} &= c_j \left(\sum_v q_{u,v} - d_j \sum_v \sum_w q_{u,w} q_{w,v} \right) = c_j \left(0 - d_j \sum_w q_{u,w} \sum_v q_{w,v} \right) \\ &= c_j(0 - d_j \times 0) = 0\end{aligned}$$

for any $d_j \in \mathbb{R}$.

For property (i) to be satisfied we need

$$l(Q) = \max\{\mathcal{L}(Q)\}, \quad \mathcal{L}(Q) = \left\{ \frac{q_{u,v}}{\sum_w q_{u,w} q_{w,v}} : u \neq v \text{ \& } \sum_w q_{u,w} q_{w,v} < 0 \right\} \quad (3)$$

and

$$u(Q) = \min\{\{\infty\} \cap \mathcal{U}(Q)\}, \quad \mathcal{U}(Q) = \left\{ \frac{q_{u,v}}{\sum_w q_{u,w} q_{w,v}} : u \neq v \text{ \& } \sum_w q_{u,w} q_{w,v} > 0 \right\}. \quad (4)$$

By definition, $l(Q) \leq 0$ and $u(Q) \geq 0$. Note that the set $\mathcal{L}(Q)$ cannot be empty, that is, $\mathcal{L}(Q) \neq \emptyset$. To prove this, suppose that $q_{a,b}$ is the largest off-diagonal element in Q . Now

$$\begin{aligned}\sum_w q_{a,w} q_{w,b} &= q_{a,a} q_{a,b} + q_{a,b} q_{b,b} + \sum_{w \neq a,b} q_{a,w} q_{w,b} \\ &= -q_{a,b} \sum_{w \neq a} q_{a,w} + q_{a,b} q_{b,b} + \sum_{w \neq a,b} q_{a,w} q_{w,b} \\ &= -q_{a,b} \sum_{w \neq a,b} q_{a,w} - q_{a,b}^2 + q_{a,b} q_{b,b} + \sum_{w \neq a,b} q_{a,w} q_{w,b}.\end{aligned}$$

However, $q_{w,b} < q_{a,b}$ for all $w \neq a$ and so

$$\sum_{w \neq a,b} q_{a,w} q_{w,b} < q_{a,b} \sum_{w \neq a,b} q_{a,w}.$$

Because $-q_{a,b}^2$ and $q_{a,b} q_{b,b}$ are strictly negative it follows that

$$\sum_w q_{a,w} q_{w,b} = -q_{a,b} \sum_{w \neq a,b} q_{a,w} + \sum_{w \neq a,b} q_{a,w} q_{w,b} - q_{a,b}^2 + q_{a,b} q_{b,b} < 0.$$

In contrast, the set $\mathcal{U}(Q)$ can be empty. Consider, for example, the rate matrix of the Jukes Cantor model, all of whose off-diagonal elements are $\delta > 0$. In this case, $\sum_w q_{u,w} q_{w,v} = -4\delta^2 < 0$ for all pairs (u, v) with $u \neq v$. Therefore $l(Q) = -1/(4\delta)$ whilst the upper limit $u(Q)$ is infinite.

To allow information to be shared between sites, we continue to assume that the coefficients $\mathbf{c} = (c_1, \dots, c_M)^T$ of the linear term are conditionally independent and identically distributed (i.i.d.) with $c_j|\alpha \sim \text{Gam}(\alpha, \alpha)$ for some unknown hyperparameter α . In an analogous fashion, we assume that the coefficients $\mathbf{d} = (d_1, \dots, d_M)^T$ of the second order term are independent of \mathbf{c} and conditionally i.i.d. with $d_j|Q, \beta \sim \mathcal{F}(\beta)$ for some unknown β , where the form of the distribution \mathcal{F} will be discussed in Section 3.2. The likelihood can then be represented as

$$p(y|Q, \tau, \ell, \alpha, \beta) = \prod_{j=1}^M \int_0^\infty \int_{l(Q)}^{u(Q)} p(c_j|\alpha) p(d_j|Q, \beta) \Pr(\mathbf{Y}_j = \mathbf{y}_j|Q_j, \tau, \ell) dc_j dd_j$$

where Q_j was defined in (2). As with the simpler LASH model, substantial gains in computational efficiency can be achieved by replacing the continuous densities $p(c_j|\alpha)$ and $p(d_j|Q, \beta)$ by discrete approximations with K_c and K_d categories, respectively. We choose to place point masses of probability $1/(K_c K_d)$ at locations $\{z_{c,a}(\alpha), z_{d,a'}(Q, \beta)\}$ for $a = 1, \dots, K_c, a' = 1, \dots, K_d$ where $z_{c,a}(\alpha)$ is the $(a-0.5)/K_c$ quantile in the distribution of $c_j|\alpha$ and $z_{d,a'}(Q, \beta)$ is the $(a'-0.5)/K_d$ quantile in the distribution of $d_j|Q, \beta$. The likelihood then simplifies to

$$p(y|Q, \tau, \ell, \alpha, \beta) \simeq \frac{1}{K_c K_d} \prod_{j=1}^M \sum_{a=1}^{K_c} \sum_{a'=1}^{K_d} \Pr[\mathbf{Y}_j = \mathbf{y}_j|Q_j \{z_{c,a}(\alpha), z_{d,a'}(Q, \beta), Q\}, \tau, \ell]. \quad (5)$$

3.1 Properties of QuASH Models

It can easily be shown that the stationary distribution of $Q_j = c_j Q - c_j d_j Q^2$ is the same as that of Q ; see Appendix A for a proof. Of course the same is also true under the simple linear scaling, $Q_j = c_j Q$, which we recover when $d_j = 0$. In the latter case, the linear mapping can simply be regarded as a site-specific scaling of the branch lengths. In contrast, our quadratic transformation does not preserve the ratios of the instantaneous rates of change in the baseline rate matrix, allowing different patterns of substitution at different sites. This idea is most readily exemplified in the context of reversible models where the transformation results in

a site-heterogeneous model in which the exchangeability parameters vary across sites. Elucidating further, it is straightforward to show that if Q is reversible, then so is Q_j ; see Appendix A for a proof. It follows that the set of GTR rate matrices is closed under our quadratic transformation. This is also true for some special cases of the GTR rate matrix including the TN93 rate matrix which was introduced in Section 2. In this case, suppose that β , ρ_1 and ρ_2 are the transversion and transition rates in the baseline rate matrix and that $\boldsymbol{\pi} = (\pi_A, \pi_G, \pi_C, \pi_T)$ is the associated stationary distribution. After applying the quadratic transformation (2), it follows from (9) in Appendix A that the transversion and transition rates in the rate matrix for site j are

$$\begin{aligned}\beta_j &= c_j \beta (1 + d_j \beta), \\ \rho_{1,j} &= c_j [\rho_1 + d_j \{\rho_1^2 - (\rho_1 - \beta)^2 \pi_R\}], \quad \rho_{2,j} = c_j [\rho_2 + d_j \{\rho_2^2 - (\rho_2 - \beta)^2 \pi_Y\}],\end{aligned}$$

where $\pi_R = \pi_A + \pi_G$ and $\pi_Y = \pi_C + \pi_T$.

As remarked in Section 2, for the simpler LASH model with all $d_j = 0$, we often fix one of the exchangeability parameters in the baseline rate matrix Q to be equal to one for parameter identifiability. For a QuASH model, scaling the baseline rate matrix Q by a constant $k > 0$ can no longer be compensated by scaling all branch lengths by $1/k$ and so an identifiability constraint is not strictly required. However, for parameter interpretability and to preserve the nested structure of the LASH and QuASH models, we continue to fix the scale of baseline rate matrix.

If we take the distribution at the root of the tree to be the vector $\boldsymbol{\pi}$ satisfying $\boldsymbol{\pi}Q = \mathbf{0}^T$ then the resulting Markov process is stationary and the term $\Pr(\mathbf{Y}_j = \mathbf{y}_j | Q_j, \tau, \boldsymbol{\ell})$ in the likelihood (5) is given by

$$\Pr(\mathbf{Y}_j = \mathbf{y}_j | Q_j, \tau, \boldsymbol{\ell}) = \sum_X \pi_{X(0)} \prod_{\text{edges } b=(v,w)} p_{j,X(v),X(w)}(\ell_b). \quad (6)$$

Here v and w are the vertices at the two ends of edge b with length ℓ_b , $X(u)$ is the character at vertex u , $u = 0$ denotes the root vertex and $P_j(\boldsymbol{\ell}) = \{p_{j,u,u'}(\ell)\}$ is the transition matrix associated with an edge of length ℓ at site j . The sum is over all functions X from the vertices to Ω_K such that $X(u)$ matches the data $y_j(u)$ for all leaf vertices u .

3.2 Random Effect Distribution

We model the coefficients $\mathbf{d} = (d_1, \dots, d_M)^T$ involved in the second order term of the quadratic transformation (2) as conditionally i.i.d. with $d_j | Q, \beta \sim \mathcal{F}(\beta)$

for some unknown hyperparameter β . As explained earlier in this section, the distribution \mathcal{F} has support on $(l(Q), u(Q))$ where $l(Q)$ is nonpositive but assumed finite whilst $u(Q)$ is nonnegative but can be infinite. This means the interval $(l(Q), u(Q))$ can be finite or semi-infinite. In order to handle the two cases in a consistent fashion, we construct the distribution of d_j through a shifted, piecewise power transformation of a Beta random variable

$$d_j = \begin{cases} l(Q) + w(Q) \left(1 - b_j^{1/w(Q)}\right), & \text{if } u(Q) \text{ is finite,} \\ l(Q) - \log b_j, & \text{otherwise,} \end{cases}$$

where $w(Q) = u(Q) - l(Q)$; $b_j|Q, \beta \sim \text{Beta}[\beta + a(Q), \beta\{b(Q) - 1\} + 1]$; and $\beta > 0$ is unknown. The terms $a(Q)$ and $b(Q)$ depend on the baseline rate matrix Q through $a(Q) = 1/w(Q)$ if $u(Q)$ is finite and $a(Q) = 0$ otherwise, and $b(Q) = \{w(Q)/u(Q)\}^{w(Q)}$ if $u(Q)$ is finite and $b(Q) = e^{-l(Q)}$ otherwise. This choice ensures that the mode of the distribution is zero, with finite probability density, and that the density of d_j decays smoothly to zero at its end points, except in the case where $l(Q) = 0$ or $u(Q) = 0$. In the special case when $l(Q) = 0$ and $u(Q)$ is infinite, the conditional distribution of d_j reduces to the $\text{Exp}(\beta)$ distribution. By centring the distribution on zero, we encourage shrinkage towards the nested LASH model with all $d_j = 0$. Although it may appear more natural to set the mean or median, rather than the mode, to zero, since the lower or upper end points of the support can be equal to zero, this is not possible in the general case.

The hyperparameter β can be assigned any prior with support on the positive real line. The dependence of the marginal prior for d_j on that for β and the parameters of the baseline rate matrix Q is complex. However, closed form expressions for the conditional expectation and variance of d_j given β , and bounds l and u can be computed and are given in Appendix B. For various values of l and u spanning the range inferred in analyses of real data, Figure 1 plots the conditional mean and standard deviation as a function of β . Clearly as β gets large, the distribution of d_j tends towards a point mass at zero and we recover a simple LASH model. However, as β approaches zero, the mean and standard deviation both become large. Therefore we can allow more heterogeneity across sites by giving β a prior which assigns reasonable density around zero.

4 Non-stationary models

The transformations characterising LASH and QuASH models allow across-site variation in the overall magnitude of the instantaneous rates of change and, for

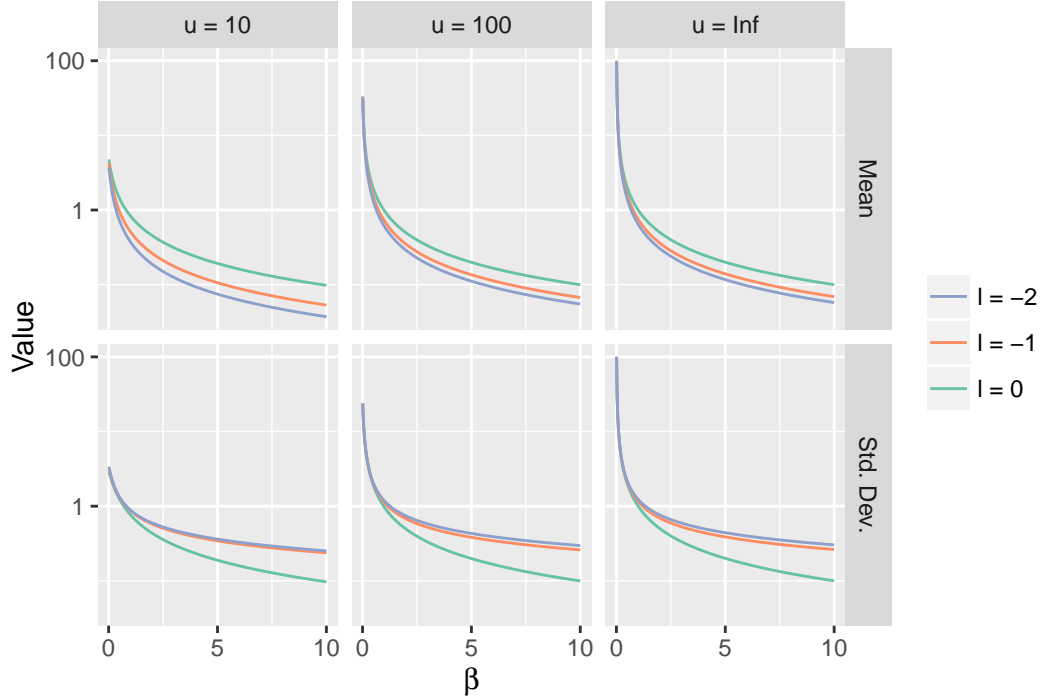


Figure 1: Conditional mean and standard deviation of d_j given β , and bounds l and u , plotted with a log-scale on the y -axis.

QuASH models, their relative sizes. However, the models discussed so far have been homogeneous across branches, with a single baseline rate matrix Q applying to the whole tree. Furthermore, the linear and quadratic transformations (1) and (2) preserve the stationary distribution π of Q . Therefore if the distribution at the root of the tree $\pi_{(0)}$ is equal to π , then the resulting Markov process will assume the same stationary distribution at all sites. These models cannot, therefore, explain the heterogeneities in sequence composition that are commonly observed in experimental data, either across taxa or across sites. As explained in Section 1, the resulting model misspecification can lead to misleading phylogenetic inferences.

Non-stationary models for sequence evolution can account for differences in composition across taxa by allowing the probability of being in each state (e.g. each nucleotide for DNA data) to change over time. Typically this is achieved by permitting step changes in the theoretical stationary distribution at different points on the tree. Although these changes do not have to occur at speciation

events (e.g. Blanquart and Lartillot 2006), this assumption is often made (e.g. Yang and Roberts 1995; Foster 2004; Heaps et al. 2014; Cherlin 2016) and we retain it here for simplicity of notation. In general, therefore, consider a rooted topology τ with B branches and a model which assumes a distribution $\pi_{(0)}$ at the root of the tree, with the processes on the other branches governed by rate matrices $Q_{(1)}, \dots, Q_{(B)}$, with associated theoretical stationary distributions $\pi_{(1)}, \dots, \pi_{(B)}$. To achieve non-stationarity we need $\pi_{(b)} \neq \pi_{(0)}$ for at least one $b \in \{1, \dots, B\}$, however for some $b \neq b'$, we might fix $\pi_{(b)}$ to be equal to $\pi_{(b')}$.

Extending the LASH and QuASH transformations to non-stationary models of this form, the rate matrix for site j on branch b is given by

$$Q_{b,j} = c_j Q_{(b)} - c_j d_j Q_{(b)}^2$$

where $c_j \in (0, \infty)$, whilst $d_j = 0$ for LASH models and $d_j \in (l, u)$ for QuASH models. In the latter case, the limits depend on all the $Q_{(b)}$, with $l = \max\{l(Q_{(b)}) : b = 1, \dots, B\}$ and $u = \min\{u(Q_{(b)}) : b = 1, \dots, B\}$, where $l(\cdot)$ and $u(\cdot)$ are as in (3) and (4) respectively. This ensures that all the resulting $Q_{b,j}$ are valid rate matrices. The likelihood expressions (5) and (6) for stationary QuASH models can now be modified to give

$$\begin{aligned} p(y|Q_{(1)}, \dots, Q_{(B)}, \pi_{(0)}, \tau, \ell, \alpha, \beta) \\ \simeq \frac{1}{K_c K_d} \prod_{j=1}^M \sum_{a=1}^{K_c} \sum_{a'=1}^{K_d} \Pr(\mathbf{Y}_j = \mathbf{y}_j | Q_{1,j}, \dots, Q_{B,j}, \pi_{(0)}, \tau, \ell) \end{aligned} \quad (7)$$

where $Q_{b,j} = Q_{b,j} \{z_{c,a}(\alpha), z_{d,a'}(Q_{(1)}, \dots, Q_{(B)}, \beta), Q_{(b)}\}$ and

$$\Pr(\mathbf{Y}_j = \mathbf{y}_j | Q_{1,j}, \dots, Q_{B,j}, \pi_{(0)}, \tau, \ell) = \sum_X \pi_{(0), X(0)} \prod_{\text{edges } b=(v,w)} p_{b,j, X(v), X(w)}(\ell_b) \quad (8)$$

in which $P_{b,j}(\ell_b) = \{p_{b,j,h,i}(\ell_b)\} = \exp(\ell_b Q_{b,j})$ is the transition matrix associated with edge b , of length ℓ_b , and site j .

By definition, non-stationary LASH and QuASH models allow heterogeneities in sequence composition across taxa. However, they also allow heterogeneities across sites. Consider, for example, a simple non-stationary model which allows a single step change in the stationary distribution at the root of the tree (e.g. Cherlin 2016, Chapter 4). In such a model, a single baseline rate matrix $Q_{(1)}$, with associated stationary distribution $\pi_{(1)} \neq \pi_{(0)}$, applies to all branches of the tree, and we denote the rate matrix associated with site j by $Q_{1,j}$. If λ is an eigenvalue

of $Q_{(1)}$, it follows immediately from (2) that $c_j\lambda - c_jd_j\lambda^2$ is an eigenvalue of $Q_{1,j}$, with $d_j = 0$ for LASH models. Denote by $\lambda_{j,1}, \lambda_{j,2}, \dots, \lambda_{j,K}$ the eigenvalues of $Q_{1,j}$ ordered such that $\lambda_{j,1} = 0 > \text{Re}(\lambda_{j,2}) \geq \text{Re}(\lambda_{j,3}) \geq \dots \geq \text{Re}(\lambda_{j,K})$, where $\text{Re}(\lambda)$ denotes the real part of the complex number λ . Under this model, it can be shown that $P_j(\ell) = \mathbf{1}\pi_{(1)} + O(e^{-\nu_j\ell})$ as $\ell \rightarrow \infty$ where $\mathbf{1}$ is a length K column vector of 1s and $\nu_j = -\text{Re}(\lambda_{j,2})$; see, for example, Kijima (1997), Chapter 4. It follows that at sites for which ν_j is large, the rate of convergence towards the stationary distribution $\pi_{(1)}$ associated with $Q_{(1)}$ will be fast, giving rise to marginal distributions at the leaves of the tree that resemble $\pi_{(1)}$. In contrast, at sites for which ν_j is small, the rate of convergence will be slow, leading to marginal distributions at the leaves that are closer to the distribution at the root $\pi_{(0)}$. Clearly these effects will vary across taxa according to the overall distance from the root to the different leaves. Although LASH and QuASH models both allow this kind of behaviour, in QuASH models it is managed more flexibly by two parameters, rather than one. Further, as discussed in Section 3.1, only the QuASH mapping allows the ratios of the instantaneous rates of change, and hence transition patterns, to vary across sites.

In the application in Section 6.2, we focus on the HB model (Heaps et al. 2014) where each branch of the tree has its own reversible rate matrix $Q_{(b)}$ which factorises into a composition vector $\pi_{(b)}$ and a set of exchangeability parameters ρ that are assumed constant across the tree. We use the formulation of the model from Williams et al. (2015) in which the composition vector on the root edge of the underlying unrooted topology is the same as that at the root of the tree $\pi_{(0)}$. To allow information to be shared between branches, the composition vectors $\{\pi_{(b)}\}$ are positively correlated *a priori*. Full details can be found in the description of Prior B in Heaps et al. (2014) but, briefly, a greater exchange of information between neighbouring branches is admitted by adopting a first order autoregressive structure in which the composition vector on branch b is conditionally independent of the composition vectors on all non-descendant branches given its parent.

5 Posterior inference via MCMC

Let θ represent the parameters of the distribution at the root of the tree and the set of baseline rate matrices. For example, $\theta = \{\pi, \rho\}$ for a simple, stationary QuASH model based on a reversible rate matrix, or $\theta = \{\pi_{(0)}, \dots, \pi_{(B-2)}, \rho\}$ for the QuASH variant of the HB model. The joint posterior distribution for all

unknowns is given by

$$\pi(\boldsymbol{\theta}, \tau, \boldsymbol{\ell}, \alpha, \beta | y) \propto p(y | \boldsymbol{\theta}, \tau, \boldsymbol{\ell}, \alpha, \beta) \pi(\boldsymbol{\theta}, \tau, \boldsymbol{\ell}, \alpha, \beta)$$

where the likelihood function $p(y | \boldsymbol{\theta}, \tau, \boldsymbol{\ell}, \alpha, \beta)$ was given in (5) and (6) for a simple, stationary QuASH model, or in (7) and (8) for a non-stationary QuASH model.

Irrespective of the choice of prior distribution $\pi(\boldsymbol{\theta}, \tau, \boldsymbol{\ell}, \alpha, \beta)$, the posterior is analytically intractable. We therefore build up a numerical approximation using a Metropolis within Gibbs sampling scheme which iterates through a series of updates for each unknown. Real valued parameters, such as branch lengths $\boldsymbol{\ell}$, can be updated using standard proposal distributions, for example Gaussian random walks on the log-scale. In QuASH models whose likelihood is invariant to the root position, τ represents an unrooted topology which can be updated using standard topological moves such as nearest neighbour interchange (NNI) and subtree prune and regraft (SPR); see, for example, Ronquist and Huelsenbeck (2003). For QuASH models whose likelihood depends on the root position, τ represents a rooted topology and so proposals which attempt to move the root are also required. In the applications in Section 6, for example, we consider the QuASH variant of the HB model and employ the NNI, SPR and root moves described in Heaps et al. (2014).

6 Applications

A controversial issue in evolutionary biology concerns the structure of the *tree of life*, whose phylogeny represents the relationships amongst its three main domains: Bacteria, Archaea and eukaryotes. There are two dominant hypotheses for the underlying unrooted topology. The classic *three domains hypothesis* of Woese et al. (1990) posits that the three domains are *monophyletic*, meaning each has an ancestor that is not shared by the others. On the basis of analyses involving previously unsequenced taxa and more sophisticated evolutionary models (Williams et al. 2013), an alternative view – the *eocyte hypothesis* – has gained considerable support over recent years. According to this conjecture, the eukaryotes emerge from within a paraphyletic Archaea, meaning the most recent common ancestor of eukaryotes and Archaea was an Archaeon. In addition to uncertainty surrounding the unrooted topology of the tree of life, opinion is also divided on the position of its root. Under the two leading hypotheses, the root is either placed on the bac-

terial branch (Gogarten et al. 1989; Iwabe et al. 1989) or, with fewer proponents, within the Bacteria (Cavalier-Smith 2006; Lake et al. 2009).

In this section we consider applications to biological data sets that address these controversial questions. In Section 6.1 we analyse a concatenated alignment of small and large subunit ribosomal RNAs (SSU and LSU rRNAs) sampled from across the tree of life. After alignment using MUSCLE (Edgar 2004) and editing to remove poorly-aligning regions, $M = 1734$ sites on $N = 36$ species remained. We consider three models: (S1) a stationary, reversible TN93 model, (S2) the LASH-variant of S1 and (S3) the QuASH-variant of S1. Models that are stationary and reversible give rise to likelihood functions that are invariant to the position of the root, and so these analyses only allow inference of the unrooted topology. In Section 6.2 we therefore consider three non-stationary models which also allow us to learn about the root position: (NS1) the HB model with TN93 exchangeability parameters; (NS2) the LASH-variant of NS1 and (NS3) the QuASH-variant of NS1. Inference via MCMC is substantially slower for the HB model and so, for computational tractability, we consider a smaller data set with $M = 1481$ sites and only $N = 16$ taxa.

In all analyses, mixing and convergence of the MCMC sampler was assessed by comparing the output from multiple chains, initialised at different starting points. In phlogenetics, mixing in tree-space can be problematic due to the low acceptance rates of topological moves. Therefore, in addition to considering the usual numerical and graphical diagnostic checks for continuous parameters, we also examined graphs based on relative cumulative split (Section 6.1) or clade (Section 6.2) frequencies of the chains over the course of the MCMC runs; see Heaps et al. (2014) for a full description of these diagnostics. Here a *split* refers to a bipartition of the taxa at the leaves of the tree into two disjoint sets, induced by cutting a branch. On a rooted tree, one of the partition subsets of any split is a *clade* if all the taxa lie on the same side of the root. In biological terms, this corresponds to an ancestor and all its descendants.

6.1 Stationary TN93 model

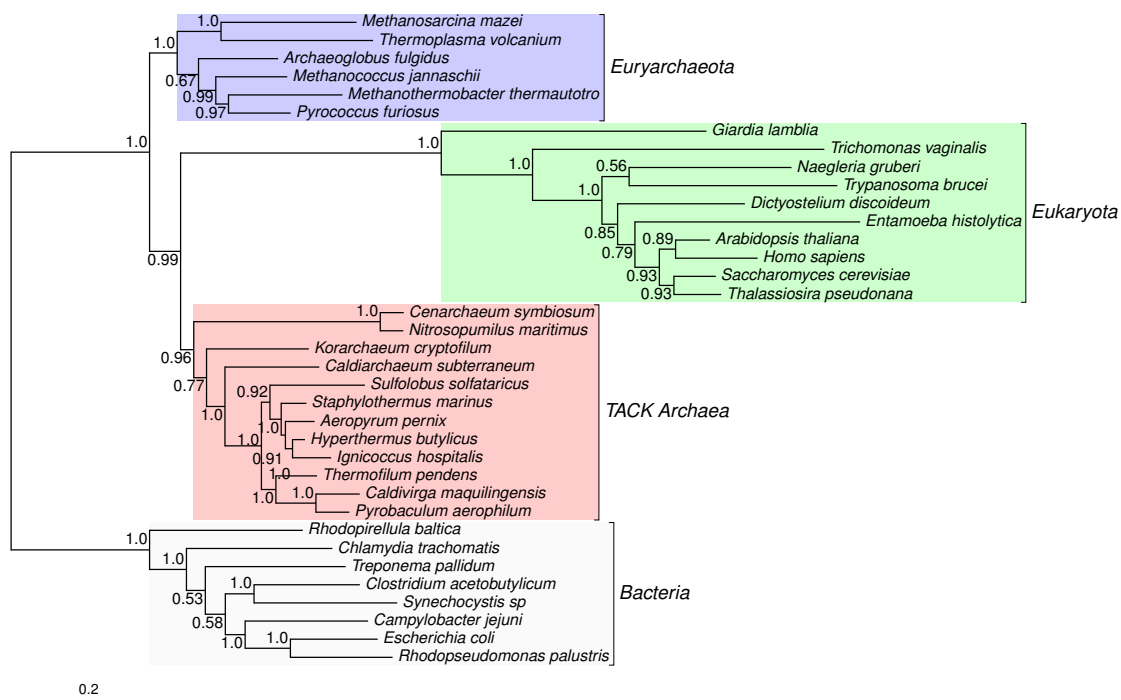
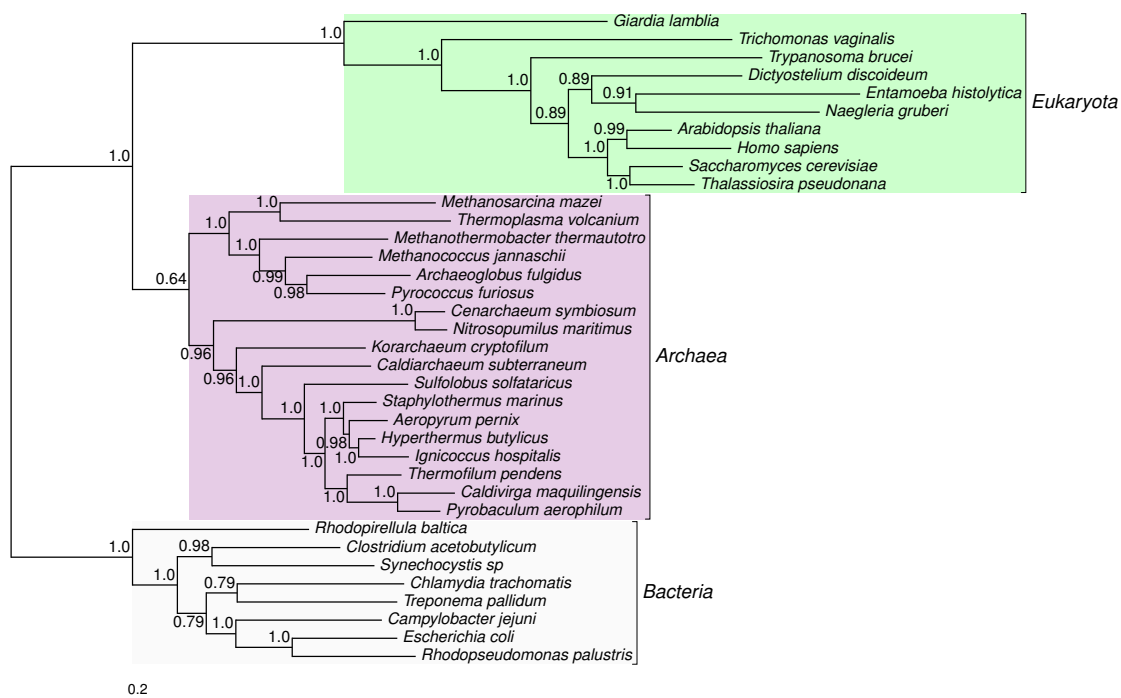
Based on our subjective assessments of the evolutionary process, for the parameters of the S1 model we chose independent gamma $\text{Gam}(1, 1)$ priors for the two transition rates ρ_1 and ρ_2 , a flat Dirichlet $\mathcal{D}(1, 1, 1, 1)$ prior for the stationary distribution π , independent exponential $\text{Exp}(10)$ priors for the branch lengths ℓ and a uniform prior over unrooted topologies τ . In models S2 and S3 we additionally assigned a gamma $\text{Gam}(10, 10)$ prior to the shape parameter α in the random ef-

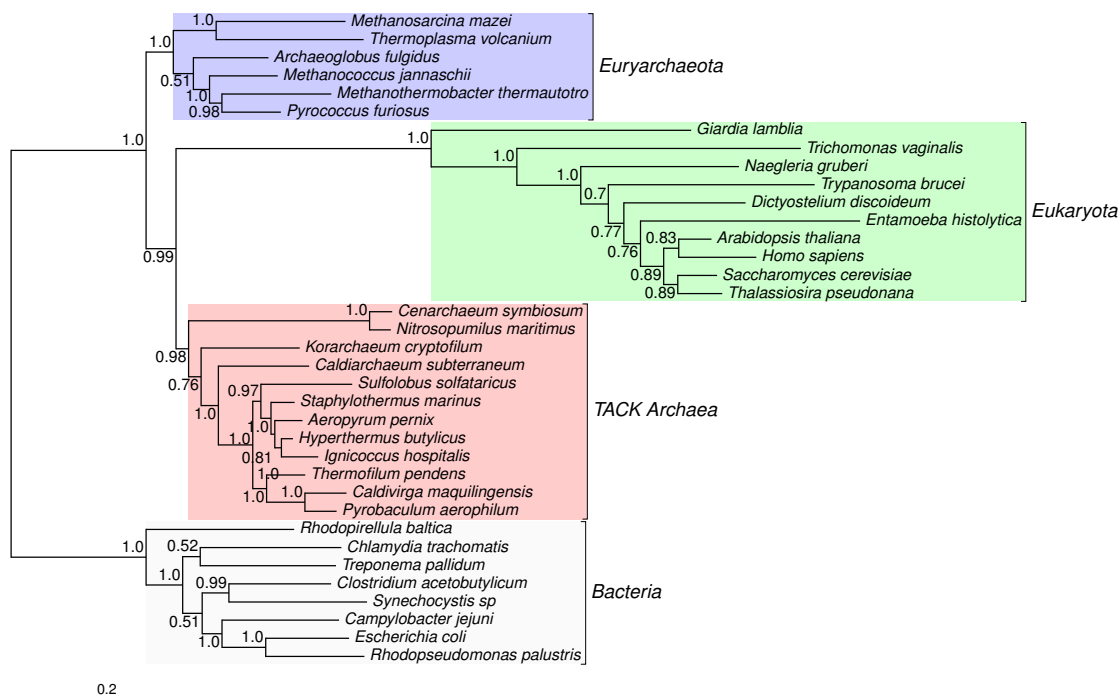
fects distribution for the rates c_j and, in model S3, a gamma $\text{Gam}(1, 1)$ prior to the parameter β in the random effects distribution for the quadratic coefficients d_j of the QuASH model. The latter distribution, with mean $E(\beta) = 1$ and coefficient of variation $CV(\beta) = 1$, was chosen to give reasonable support to values of β near zero. As explained in Section 3.2, this choice makes the prior for the d_j reasonably diffuse. In order to check sensitivity to the prior specification for β , we repeated the analysis with model S3 using priors that had the same mean but different coefficients of variation and different behaviour near zero: $\text{Gam}(10, 10)$ ($CV(\beta) = 0.316$) and $\text{Gam}(0.1, 0.1)$ ($CV(\beta) = 3.16$). The phylogenetic and posterior predictive inferences reported in this section were robust against these changes.

For each model the MCMC algorithm outlined in Section 5 was used to generate at least $110K$ draws from the posterior, after a burn-in of $100K$ samples, thinning the remaining output to retain every 100th iterate. The diagnostics checks described earlier gave no evidence of any lack of convergence.

In phylogenetic inference, the majority-rule consensus tree is the most widely used summary of the posterior distribution over tree space. As a summary of a sample of trees, it includes only those splits which appear in over half of the samples (Bryant 2003), here representing those with posterior probability greater than 0.5. For the analyses under models S1, S2 and S3, the consensus trees are shown in Figure 2 in which the numerical labels represent the posterior probability of the associated split. To aid comparison, the trees are all visualised with the root at the midpoint of the bacterial branch. The consensus tree under S1 supports the three domains hypothesis, whilst models S2 and S3 yield eocyte trees, with eukaryotes emerging from within two archaeal clades: the Euryarchaeota and the TACK Archaea. As expected, there is a marked difference in our phylogenetic inferences as we move from the simple TN93 model (S1) to one which incorporates across-site rate heterogeneity. However, there is very little difference in the inferences obtained when extending the LASH model (S2) to the corresponding QuASH model (S3). Comparing the prior and posterior density for β in Figure 3, the posterior seems to support larger values for β than the prior, which suggests a distribution for the quadratic coefficients d_j that is more concentrated around zero. The data do not, therefore, provide much evidence that the QuASH transformation is necessary given a model that already incorporates across-site rate heterogeneity.

As explained in Section 1, functional and structural constraints acting on a particular site can cause it to evolve very slowly. In such cases we are likely to see little or no variation in the character state at that column of the alignment. Therefore in fitting to the alignment-wide empirical compositions, models that do





(c)

Figure 2: Majority rule consensus trees under models (a) S1, (b) S2 and (c) S3. Numerical labels represent the posterior probability of the associated split, and branch lengths can be interpreted as the expected number of substitutions per site. Trees are unrooted but visualised with the root at the midpoint of the bacterial branch.

not allow variation in, at least, the rate of the evolutionary process across sites tend to overestimate the mean number of distinct nucleotides per column, and underestimate the associated standard deviation. Figure 4 shows the posterior predictive distribution for these test statistics obtained under models S1, S2 and S3, together with the observed values calculated from the alignment. As expected, model S1 markedly overestimates the number of distinct nucleotides per site and underestimates the associated standard deviation. Whilst models S2 and S3 also overestimate the mean, the discrepancies are much less marked, with the QuASH-variant of the TN93 model (S3) being most compatible with the observed data. Interestingly, models S2 and S3 overestimate the standard deviation of the number of distinct nucleotides per site, slightly more noticeably for model S2 than S3. It

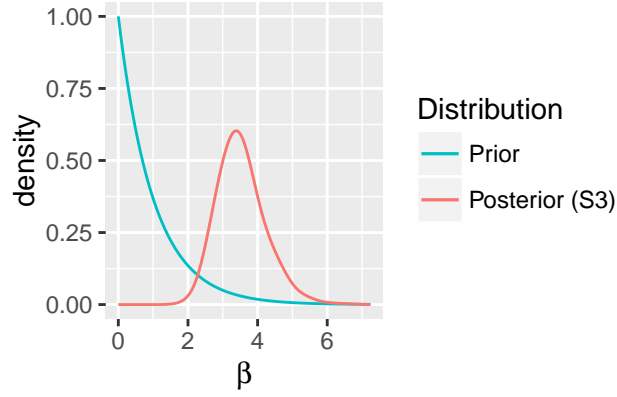


Figure 3: Marginal prior and posterior densities for the unknown parameter β in the random effect distribution for the quadratic coefficients d_j .

is possible that models allowing sequence composition to vary across sites would be required to adequately capture this feature.

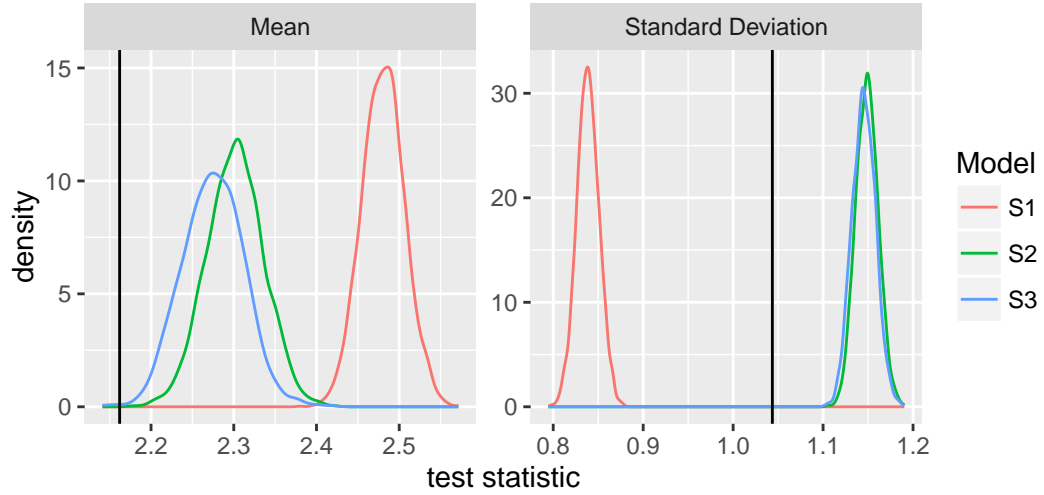


Figure 4: Posterior predictive densities for the mean and standard deviation of the number of distinct nucleotides per site. The observed values are indicated by vertical lines.

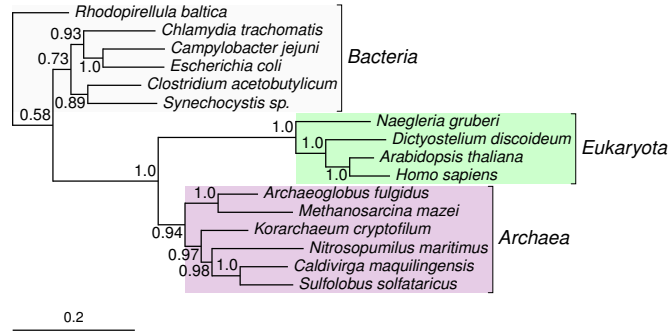
6.2 Non-stationary HB model

For the analyses using the non-stationary models NS1, NS2 and NS3, we adopted the prior distributions outlined in Section 6.1 for the two transition rates ρ_1 and ρ_2 , the branch lengths ℓ and the parameters α and β in the random effects distributions for the linear and quadratic coefficients c_j and d_j . As the HB model yields a likelihood function that depends on the position of the root, our topology τ is rooted. We assigned τ a prior according to the biologically-motivated Yule model of speciation, which generates a distribution in which near equal probability is assigned to root splits of all sizes: $1:N-1$, $2:N-2$, and so on (Cherlin et al. 2015). For the composition vectors π_b , $b = 0, \dots, B-2$, in the baseline rate matrix we used Prior B from Heaps et al. (2014), choosing the hyperparameters representing the autoregressive coefficient and conditional variance to be $a = 0.94$ and $b = 0.31$ respectively. This specification was guided by simulations from the prior predictive distribution which suggested it led to a biologically plausible degree of heterogeneity in empirical sequence composition.

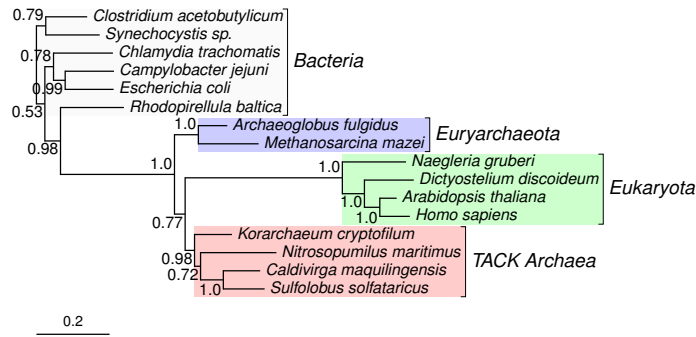
For each model the MCMC algorithm was used to generate at least $510K$ draws from the posterior, after a burn-in of $500K$ samples, thinning the remaining output to retain every 100th iterate. The diagnostics checks described earlier gave no evidence of any lack of convergence.

The rooted majority-rule consensus trees for each model are shown in Figure 5. Our conclusions are consistent with those from Section 6.1. Specifically, the model NS1 supports a three-domains tree whilst models NS2 and NS3 support very similar eocyte trees with, in this case, the same rooted topology. Although the site-homogeneous HB model (NS1) and the LASH and QuASH variants (NS2 and NS3) support different conclusions with regards to the unrooted topology, they both suggest a root within the Bacteria. The marginal posterior distribution for root splits under the three models is summarised in Table 1. Again, the differences between the inferences under NS1 and NS2 are much more marked than the differences between those under NS2 and NS3. However, in all cases the posterior probability for a root within the Bacteria is 1.0.

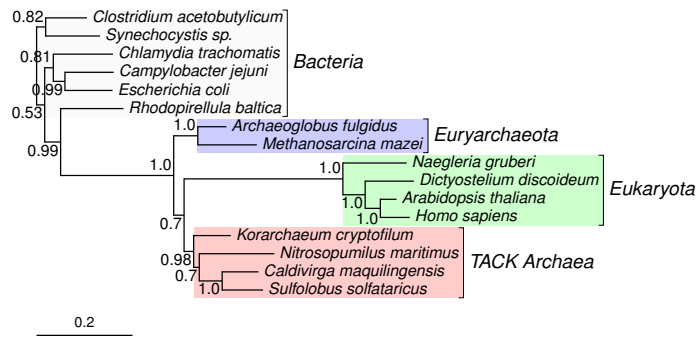
The LASH and QuASH variants of the HB model allow sequence composition, as well as the overall rate of evolution, to vary across sites. Therefore we expect these models to be better equipped to capture the number of distinct nucleotides per site. Posterior predictive densities of the across-site mean and standard deviation are plotted in Figure 6. For the mean, all three models capture the observed statistic well, with the site-homogeneous model (NS1) offering slightly more support to larger values, as expected. As in the analysis from Sec-



(a)



(b)



(c)

Figure 5: Rooted majority rule consensus trees under models (a) NS1, (b) NS2 and (c) NS3. Numerical labels represent the posterior probability of the associated clade, and branch lengths can be interpreted as the expected number of substitutions per site.

Table 1: Root splits receiving posterior support of at least 0.01 under the NS1, NS2 or NS3 models, and the associated probabilities. The root split is the bipartition induced by cutting the branch edge of the associated unrooted topology; the smaller partition is shown.

Root Split	Posterior Probability		
	NS1	NS2	NS3
<i>Rhodopirellula baltica</i>	0.576	0.017	0.015
<i>Campylobacter jejuni</i> , <i>Chlamydia trachomatis</i> , <i>Clostridium acetobutylicum</i> , <i>Escherichia coli</i> , <i>Synechocystis sp. PCC6803</i>	0.153	0.067	0.088
<i>Clostridium acetobutylicum</i> , <i>Synechocystis sp. PCC6803</i>	0.085	0.324	0.343
<i>Synechocystis sp. PCC6803</i>	0.108	0.197	0.172
<i>Campylobacter jejuni</i> , <i>Chlamydia trachomatis</i> , <i>Escherichia coli</i>	0.040	0.176	0.189
<i>Campylobacter jejuni</i> , <i>Escherichia coli</i>	0.031	0.176	0.157
<i>Campylobacter jejuni</i> , <i>Clostridium acetobutylicum</i> , <i>Escherichia coli</i> , <i>Synechocystis sp. PCC6803</i>	0.004	0.018	0.014
<i>Clostridium acetobutylicum</i>	0.001	0.012	0.012

tion 6.1, the site-homogeneous model very markedly underestimates the standard deviation. The posterior predictive densities under the LASH (NS2) and QuASH (NS3) variants of the HB model are very similar. Although both overestimate the standard deviation, the observed statistic is more plausible than under the NS1 model, and the overestimation seems less marked than the corresponding comparison from Section 6.1. The similarity in both phylogenetic and posterior predictive inferences under the LASH and QuASH models are consistent with the implications of the comparison between the prior and posterior for β in Figure 7. Again, the posterior seems to support larger values of β than the prior which suggests

that the QuASH transformation adds little given a model in which linear across-site heterogeneity is already included.

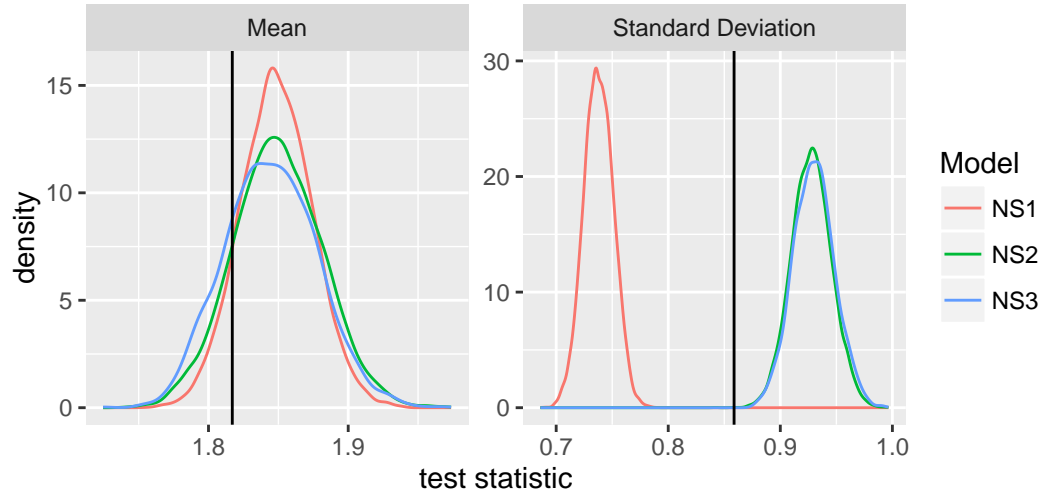


Figure 6: Posterior predictive densities for the mean and standard deviation of the number of distinct nucleotides per site. The observed values are indicated by vertical lines.

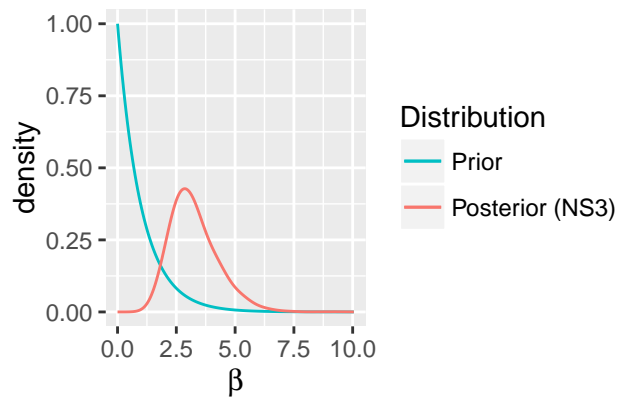


Figure 7: Marginal prior and posterior densities for the unknown parameter β in the random effect distribution for the quadratic coefficients d_j .

7 Discussion

The introduction of across-site rate heterogeneity into substitution models for sequence evolution led to substantial improvements in model fit and the credibility of phylogenetic inferences. In practice, this feature was incorporated through a set of site-specific rates, modelled as random effects with a unit mean gamma distribution, that linearly transformed a baseline rate matrix. Motivated by the advancement gained through this simple innovation, we considered a natural extension based on the incorporation of two sets of random effects, allowing site-specific quadratic transformation of the baseline rate matrix. We derived properties of QuASH-transformed rate matrices, showing that they retain the stationary distribution of the underlying baseline matrix, and that the set of reversible rate matrices is closed under our quadratic transformation. In the context of a class of non-stationary models which permit step-changes in the theoretical stationary distribution at one or more points on the tree, we demonstrated that both the LASH and QuASH transformations lead to models which allow sequence composition to vary across sites as well as across taxa. This is due to different rates of convergence towards the theoretical stationary distributions at different sites. The QuASH-transformed, non-stationary models therefore provide a parsimonious means of allowing heterogeneity in sequence composition across both alignment dimensions.

We utilised our model and inferential procedures in two biological applications concerning the tree of life. In the first, we compared inferences under a stationary, reversible TN93 model, with those obtained under the LASH and QuASH extensions. In the second, to make computational inference manageable, we considered a smaller data set and compared inferences under a non-stationary HB model to those obtained under the LASH and QuASH variants. In both applications we found that the simpler site-homogeneous models supported the three domains hypothesis, with the Archaea, Bacteria and eukaryotes appearing as monophyletic groups. Conversely the more flexible LASH and QuASH models supported the eocyte hypothesis, with eukaryotes emerging from within a paraphyletic Archaea. The non-stationary models consistently supported a root within the Bacteria. The marked differences between inferences obtained under the site-homogeneous and LASH models are congruent with other results reported in the literature (Yang 1996). However, neither analysis suggested that the quadratic transformation added much value once a linear transformation was in place. We have drawn similar conclusions from applications to several other data sets not reported here.

Although our analyses have reinforced the importance of allowing heterogene-

ity in the rate of evolution across sites, it appears that the natural extension, exploiting a quadratic transformation of the base rate matrix, adds little value. However, in the context of non-stationary models, it is worth emphasising that even the LASH transformation generates models that allow heterogeneity in sequence composition across sites as well as across taxa. To our knowledge, this is a property that has gone unnoticed in the literature. Whilst a few, more mechanistic models have been proposed to offer this flexibility (e.g. Blanquart and Lartillot 2008; Jayaswal et al. 2014), their complexity has made model-fitting computationally prohibitive. In contrast, non-stationary LASH and QuASH models provide a more parsimonious, data-driven alternative for which computational inference is substantially more straightforward.

Acknowledgements

S.E.H., T.A.W., S.C. and T.M.E. were supported by a European Research Council Advanced Investigator Award (ERC-2010-AdG-268701), and a Program Grant from the Wellcome Trust (number 045404) to T.M.E.

References

- S. Blanquart and N. Lartillot. A Bayesian compound stochastic process for modeling non-stationary and nonhomogeneous sequence evolution. *Molecular Biology and Evolution*, 23:2058–2071, 2006.
- S. Blanquart and N. Lartillot. A site- and time-heterogeneous model of amino-acid replacement. *Molecular Biology and Evolution*, 25:842–858, 2008.
- D. Bryant. A classification of consensus methods for phylogenies. In M. Janowitz, F.-J. Lapointe, F. R. McMorris, B. Mirkin, and F. S. Roberts, editors, *Bioconsensus, DIMACS Series*, pages 163–184, Providence, Rhode Island, 2003. American Mathematical Society.
- T. Cavalier-Smith. Rooting the tree of life by transition analyses. *Biology Direct*, 1(1):1–83, 2006.
- S. Cherlin. *Rooting major cellular radiations using statistical phylogenetics*. PhD thesis, Newcastle University, 2016.

- S. Cherlin, T. M. W. Nye, S. E. Heaps, R. J. Boys, T. A. Williams, and T. M. Embley. Investigating the effect of non-reversibility on inferring rooted phylogenies. Available at arXiv:1505.08009, 2015.
- C. J. Cox, P. G. Foster, R. P. Hirt, S. R. Harris, and T. M. Embley. The archaeobacterial origin of eukaryotes. *Proceedings of the National Academy of Sciences*, 105(51):20356–20361, 2008.
- R. C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797, 2004.
- T. M. Embley, R. H. Thomas, and R. A. D. Williams. Reduced thermophilic bias in the 16S rDNA sequence from thermus ruber provides further support for a relationship between thermus and deinococcus. *Systematic and Applied Microbiology*, 16(1):25–29, 1993.
- W. M. Fitch and E. Markowitz. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochemical Genetics*, 4(5):579–593, 1970.
- P. G. Foster. Modeling compositional heterogeneity. *Systematic Biology*, 53:485–495, 2004.
- N. Galtier and J. R. Lobry. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *Journal of Molecular Evolution*, 44(6):632–636, 1997.
- J. P. Gogarten, H. Kibak, P. Dittrich, L. Taiz, E. J. Bowman, B. J. Bowman, M. F. Manolson, R. J. Poole, T. Date, and T. Oshima. Evolution of the vacuolar H⁺-ATPase: implications for the origin of eukaryotes. *Proceedings of the National Academy of Sciences*, 86(17):6661–6665, 1989.
- S. E. Heaps, T. M. W. Nye, R. J. Boys, T. A. Williams, and T. M. Embley. Bayesian modeling of compositional heterogeneity in molecular phylogenetics. *Statistical Applications in Genetics and Molecular Biology*, 13(5):589–609, 2014.
- N. Iwabe, K. Kuma, M. Hasegawa, S. Osawa, and T. Miyata. Evolutionary relationship of archaeobacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proceedings of the National Academy of Sciences*, 86(23):9355–9359, 1989.

- V. Jayaswal, T. K. F. Wong, J. Robinson, L. Poladian, and L. S. Jermin. Mixture models of nucleotide sequence evolution that account for heterogeneity in the substitution process across sites and across lineages. *Systematic Biology*, 63(5): 726–742, 2014.
- M. Kijima. *Markov processes for stochastic modeling*. Chapman & Hall, London, first edition, 1997.
- J. A. Lake, R. G Skophammer, C. W. Herbold, and J. A. Servin. Genome beginnings: rooting the tree of life. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1527):2177–2185, 2009.
- N. Lartillot, H. Brinkmann, and H. Philippe. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evolutionary Biology*, 7(1):1–14, 2007.
- H. Philippe, H. Brinkmann, D. V. Lavrov, D. T. J. Littlewood, M. Manuel, G. Wörheide, and D. Baurain. Resolving difficult phylogenetic questions: Why more sequences are not enough. *PLoS Biology*, 9(3):1–10, 2011.
- F. Ronquist and J. P. Huelsenbeck. MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19:1572–1574, 2003.
- N. Sueoka. Directional mutation pressure and neutral molecular evolution. *Proceedings of the National Academy of Sciences*, 85(8):2653–2657, 1988.
- T. Uzzell and K. W. Zorbin. Fitting discrete probability distributions to evolutionary events. *Science*, 172(3988):1089–96, 1971.
- T. A. Williams, P. G. Foster, C. J. Cox, and T. M. Embley. An archaeal origin of eukaryotes supports only two primary domains of life. *Nature*, 504:231–236, 2013.
- T. A. Williams, S. E. Heaps, S. Cherlin, T. M. W. Nye, R. J. Boys, and T. .M. Embley. New substitution models for rooting phylogenetic trees. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1678), 2015.
- C. R. Woese, O. Kandler, and M. L. Wheelis. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences*, 87(12):4576–4579, 1990.

- Z. Yang. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution*, 10 (6):1396–1401, 1993.
- Z. Yang. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution*, 39(3):306–314, 1994.
- Z. Yang. Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology & Evolution*, 11(9):367–372, 1996.
- Z. Yang and D. Roberts. On the use of nucleic acid sequences to infer early branchings in the tree of life. *Molecular Biology and Evolution*, 12:451–458, 1995.

Appendix A: Proofs of Properties of QuASH Models

Stationary Distribution

It is easy to show that the stationary distribution of $Q_j = c_j Q - c_j d_j Q^2$ is the same as that of Q . Suppose the row vector π is the stationary distribution of Q . Then $\pi Q = \mathbf{0}^T$ where $\mathbf{0}$ is a column vectors of 0s. It follows that

$$\pi Q_j = \pi(c_j Q - c_j d_j Q^2) = c_j(\pi Q) - c_j d_j(\pi Q)Q = \mathbf{0}^T.$$

Hence π is the stationary distribution of Q_j .

Reversibility

Suppose that the base matrix Q is reversible. We can therefore write $Q = S\Pi$ where S is a symmetric matrix of exchangeability parameters and $\Pi = \text{diag}(\pi)$. Since the stationary distribution of Q_j is also π , Q_j is reversible if we can find a symmetric matrix S_j such that $Q_j = S_j\Pi$. Now $Q_j = c_j Q - c_j d_j Q^2 = c_j S\Pi - c_j d_j S\Pi S\Pi = c_j(S - d_j S\Pi S)\Pi$ and so we need to check whether the matrix $(S - d_j S\Pi S)$ is symmetric. The difference of two symmetric matrices is symmetric and so this is tantamount to checking whether $S\Pi S$ is symmetric.

The matrix S is symmetric and we denote the (u, v) -th entry by $s_{\min(u,v), \max(u,v)}$. The matrix $S\Pi$ has (u, v) -th entry $s_{\min(u,v), \max(u,v)}\pi_v$ and so the corresponding element of $S\Pi S$ is given by

$$(S\Pi S)_{u,v} = \sum_w (S\Pi)_{u,w} s_{\min(w,v), \max(w,v)} = \sum_w s_{\min(u,w), \max(u,w)} \pi_w s_{\min(w,v), \max(w,v)}.$$

The expression on the right-hand-side is exchangeable with respect to the indices u and v and so $(S\Pi S)_{u,v} = (S\Pi S)_{v,u}$. Therefore the matrix $(S - d_j S\Pi S)$ is symmetric and so Q_j is the rate matrix of a reversible model with exchangeability matrix

$$S_j = c_j(S - d_j S\Pi S). \quad (9)$$

Appendix B: Conditional Expectation and Variance of $(d_j|\beta, l, u)$

It is straightforward to show that the conditional expectation and variance of the quadratic coefficient d_j given β , lower limit l and (finite) upper limit u are given by

$$E(d_j|\beta, l, u) = l + w \left[1 - \frac{\Gamma(\beta + 2/w)\Gamma\{\beta(w/u)^w + 1/w + 1\}}{\Gamma(\beta + 1/w)\Gamma\{\beta(w/u)^w + 2/w + 1\}} \right]$$

and

$$\text{Var}(d_j|\beta, l, u) = w^2 \left[\frac{\Gamma(\beta + 3/w)\Gamma\{\beta(w/u)^w + 1/w + 1\}}{\Gamma(\beta + 1/w)\Gamma\{\beta(w/u)^w + 3/w + 1\}} - \frac{\Gamma(\beta + 3/w)^2\Gamma\{\beta(w/u)^w + 1/w + 1\}^2}{\Gamma(\beta + 1/w)^2\Gamma\{\beta(w/u)^w + 3/w + 1\}^2} \right]$$

where $w = (u - l)$.

Similarly, when the upper limit u is infinite, the corresponding expressions are given by

$$E(d_j|\beta, l) = l - \{\Psi(\beta) - \Psi(\beta e^{-l} + 1)\} \quad \text{and} \quad \text{Var}(d_j|\beta, l) = \Psi_1(\beta) - \Psi_1(\beta e^{-l} + 1),$$

where $\Psi(\beta) = d \ln \Gamma(\beta) / d\beta$ and $\Psi_1(\beta) = d^2 \ln \Gamma(\beta) / d\beta^2$ are the digamma and trigamma functions.